N P R S T

Navy Personnel Research, Studies, and Technology
5720 Integrity Drive • Millington, Tennessee 38055-1000 • www.nprst.navy.mil

research at work

A Study of Alternative Modeling Techniques of Attrition of First-term Navy Enlisted Sailors

J. Scott Johnson, **Ph.D.**
Jerry C. Crabb, **M.A.**

A Study of Alternative Modeling Techniques of Attrition for First-Term Navy Enlisted Sailors

J. Scott Johnson, Ph.D.
Jerry C. Crabb, M.A.

Reviewed and Approved by
David M. Cashbaugh
Institute for Force Management Sciences

Paul Rosenfeld, Ph.D.
Institute for Organizational Assessment

Released by
David L. Alderton, Ph.D.
Director

REPORT DOCUMENTATION PAGE

1. REPORT DATE *(DD-MM-YYYY)* 31-05-2009	2. REPORT TYPE Technical Note	3. DATES COVERED *(From - To)* 7 July 08- 14 August 08

4. TITLE AND SUBTITLE

A Study of Alternative Modeling Techniques of Attrition of First-term Navy Enlisted Sailors

5a. CONTRACT NUMBER

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

J. Scott Johnson, Ph.D.
Jerry C. Crabb, M.A.

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

Navy Personnel Research, Studies, and Technology (NPRST/BUPERS-1)
Bureau of Naval Personnel
5720 Integrity Dr.
Millington, TN 38055-1000

8. PERFORMING ORGANIZATION REPORT NUMBER

NPRST-TN-09-7

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

Office of Naval Research
800 North Quincy Street
Ballston Tower One
Arlington, VA 22217-5660

10. SPONSOR/MONITOR'S ACRONYM(S)

ONR

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT

A - Approved for public release; distribution is unlimited.

13. SUPPLEMENTARY NOTES

14. ABSTRACT

The purpose of this study was to examine enlisted Navy first-term attrition. Sailors who exit the Navy before completing the end of their first term of enlistment are a lost investment of direct costs (training) and indirect costs (force stability). The ability to predict individuals that are likely to attrite, and to accurately forecast the percentage of the force that is likely to leave early can potentially allow the Navy to proactively implement retention strategies. These strategies will result in maximization of training dollars invested in top performing Sailors.
This study uses modeling techniques and conditional probabilities in an attempt to generate an improved predictive ability of attrition.

15. SUBJECT TERMS

Attrition, Data Mining, First Term-Attrition

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE	UNCLASS	27	Genni Arledge
UNCLASS	UNCLASS	UNCLASS			19b. TELEPHONE NUMBER *(Include area code)* 901-874-2115 (882)

Foreword

The focus of this research project was to analyze probabilities of first-term attrition for enlisted Sailors from the U.S. Navy by employing advanced statistical techniques in the domain of data mining. The research was conducted as an initial effort for the Taxonomy of Self-Reports (TAXSE, Data Mining) project under Personnel Integration of Selection, Classification, Evaluations, and Surveys (PISCES) sponsored by the Office of Naval Research (ONR). The outcome of the research would then be compared to the analysis performed and results derived by Dr. Amos Golan, a leading econometrician, who utilized classical econometric techniques to predict the overall rate attrition for a segment of enlisted Sailors.

The authors would also like to thank ONR for their generous sponsorship.

David L. Alderton, Ph.D.
Director

Contents

Introduction

Objective

Sailors who separate from the Navy before completing the end of their first term of enlistment are a lost investment of direct costs (training) and indirect costs (force stability). The ability to predict individuals who are likely to attrite, and to accurately forecast the percentage of the force that is likely to separate prior to completing their obligation allows the Navy to proactively implement retention strategies. These strategies will result in maximization of training dollars invested in top performing Sailors.

The objective of this study is to generate a predictive model of first-term enlisted attrition using the advanced statistical techniques of data mining. This would allow the Navy to predict first-term enlisted attrition by looking at individual characteristics. It can then be used with conditional probabilities to estimate total Navy attrition of first-term Sailors.

Background

There have been numerous attempts to predict attrition. The United States Army Research Institute attempted to predict some of the most important instrumental variables of attrition rates (Salo & Siebold, 2006). In their study, they examined Finnish Army Conscripts in an attempt to generalize their models. They considered aptitudes, mental and physical health, and attitudes. The researchers accomplished this by administering questionnaires at three different points in time and using Likert scales for each variable. The investigators also employed standard log-linear regression as well as Cox-regression in their analysis. Because of the data and techniques that were used, Salo and Siebold (2006) were only able to explain 30 to 40 percent of the variance with large misclassification rates.

A study performed by the Naval Personnel Research, Studies, and Technology, First Watch on the First Term on Enlistment, White and Mottern (2007) was a longitudinal project examining the psychology behind attrition. This project tracked the first-term Sailor from initial entry through "A" School and through the first year in the fleet. Questionnaires were administered at different discrete periods throughout this time period. Some areas addressed by the survey included demographics, reasons for joining the Navy, experiences in training, and social support structures while in the training pipeline. Some of the findings showed that people who tend to attrite initially join the Navy to escape from some type of home situation, were generally more negative about the possibility of a career in the Navy, and lack of motivation or boredom.

In January 2008, the Center for Naval Analysses released a study of first-term attrition, Attrition and Reenlistment of First Term Sailors through FY07. Some of the key findings in this study were that females generally had a higher attrition rate than men. Also of significance was the fact that Sailors with waivers experienced greater rates of attrition, B cell Sailors with respect to AFQT scores tended to attrite more while A Cell sailors, those that recruiters go after, had the lowest rates of attrition.

Buttrey and Larson (1999) conducted an analysis for the U. S. Army and Office of the Deputy Chief of Staff, Personnel (ODCSPER). The authors used Classification and Regression Tree (CART) algorithms. This method allowed grouping individuals based on individual characteristics to determine attrition probability. Buttrey and Larson devised new methods to group individuals entering military service in order to predict attrition. They also reduced the misclassification rate found in previous studies. They approached the issue of classification based on characteristics differently from Salo and Siebold by using a different mathematical algorithm known as the Classification and Regression Tree (CART).

An Air Force technical report, Flyer (1959) examined first-term enlistment using pre-enlistment indicators such as high school diplomas. McCloy, DiFazio, and Carter (1993) extended this idea by examining non-cognitive pre-enlistment indicators. It was posited that some of these would directly impact military job performance (discipline, physical fitness, e.g.) as well as impacting attrition. McCloy et al. studied biodata, temperament data, interest measures, and measures of job preferences. They then attempted to predict attrition using Proportional Hazard analysis (Cox, 1972). The authors found that high school graduation was a strong predictor. Non-high school graduates had a hazard rate for attrition of approximately twice that of graduates. They also found that non-cognitive pre-enlistment biodata significantly increased the power to predict attrition more than any other explanatory variables. This study adequately predicted individual attrition. However, it lacked the ability to be scaled up to predict the percentage of populations in the military that might attrite.

Discussion

In forecasting military manpower, it is extremely valuable to know what proportion of the force is expected to leave abruptly. However, it will be highly beneficial to the Navy to know the individual characteristics that are likely to lead to attrition. This could lead to cost savings to the Navy in high attrition ratings such as SEALS, where significant training dollars are invested in individual Sailors with very valuable skill sets. The current model is different in this regard and more valuable, and applicable in many aspects. It can predict attrition for both individuals and for entire military populations.

Data Considerations

The hypothesis of this work was that one could more accurately predict attrition using alternative data mining methods than those that have been reported in the past literature. Most of the past work used standard econometric approaches to modeling attrition. A few of the previous studies used some decision tree methods. However, the authors believed that previous research did not capture all of the information sufficiently.

The data used in this research project came from another attrition research project conducted for the Institute of Force Management Sciences by Dr. Amos Golan. Using a data set spanning from 2001 through 2008, his analysis focused specifically on the U.S. Navy's occupational group "Aviation Ground Support." An important consideration with this data was the fact that a Sailor could only be selected for inclusion by having a specific career event occur. An "event" was defined as one of four possible decision points: reenlistment, extension, separation (leaving the Navy due to EAOS), or attrition.

Estimation Methods

Data mining has gained tremendous popularity in its usage within the private sector during the last decade. The banking industry uses predictive modeling for fraud detection and scoring customers who are the best and worst risks for marketing offerings. Direct marketers utilize similar techniques for a variety of purposes. Consumers that are identified as most likely to respond to certain marketing offers add profitability to the company. Conversely, in the case of companies sending out mail solicitations to millions of consumers, ideally the companies want to know the point where profitability and loss behavior occurs in order understand where to make the cutoff of customer segments with which to mail. This research is the first application of data mining techniques to address Navy Manpower and Personnel Research issues.

The software used for this project was Clementine developed by SPSS. This data mining software follows the Cross-Industry Standard Process for Data Mining (CRISP-DM) approach. This approach involves five distinct phases of business understanding: Data Understanding, Data Preparation, Modeling, Evaluating, and eventual Deployment of the model.

Since the dependent variable "attrite" in this model was binary, algorithms commensurate with binary classification modeling were utilized. These techniques such as neural networks, logistic regressions, and decision tree algorithms were simultaneously ran to obtain an optimal solution. Rule sets were specified in order to discard any model that did not give at least 80 percent predictive power (area under the curve or overall accuracy) or better.

Analysis

The best models to predict attrition were the Chi Square Automatic Interaction Detection (CHAID) (decision tree) and Logistic Regression. The models were then compared by looking at lift charts in Appendix A. Lift charts put the observations in the validation set into increasing or decreasing order on the basis of the score. The score is the probability of response event (success), which is estimated on the training data. These are then subdivided into deciles and then the graph is calculated. A model is said to be valid if the observed success probabilities follow the same order (increasing or decreasing) as the estimated probabilities. A lift chart is usually compared with a baseline curve. The baseline is interpreted as the probability estimates in the absence of the model. Another way to think of this is taking the mean of the observed success probabilities. By dividing the values of each curve by the baseline, a relative index of performance, the lift, is obtained.

The CHAID algorithm performed the best with 94 percent predictive power in both the training and testing partitions. CHAID works by analyzing the Chi square statistic for each split for the independent variables of interest. Tree models can be thought of as a recursive procedure through which a set of statistical units are progressively divided into groups. This is accomplished by a division rule that strives to maximize homogeneity of the response variable in each of the groups. At each step of the procedure a division rule is specified by the choice of an explanatory variable to split and the choice of a splitting rule for variables which establishes how to partition variables.

The main result of a tree is a partition of variables. The impurity measure (i.e., variance) used by the CHAID methods is the distance between the observed and expected frequencies. The expected frequencies are calculated using the hypothesis for homogeneity for the observations in the considered node. This split criterion function is the Pearson X^2 index. If the decrease in X^2 is significant (p value is less than the pre-specified α) then the node is split. If not, it remains unsplit and becomes a leaf. The CHAID stops tree growth through a stopping criteria based on the X^2 test. An advantage of the CHAID model is that it produces a wider tree. This is advantageous because there are more splits on categories and ranges of categories such as length of service instead of only one category. This is a standard methodology in marketing segmentation studies.

Construction of model ensembles or meta-models is one of the most powerful techniques used in predictive modeling. This takes the best of the individual models and combines the equations for each into a larger, and more robust predictive model. This approach was tested by combining the CHAID and Logistic regression models. However, after again examining the lift curves and overall accuracy, the CHAID model was chosen to be the best model in terms of accurately predicting first-term Sailor attrition. Variables that were found to be predictive included annual housing and subsistence allowance, base pay, dependent status, sea or shore duty, unemployment rates, S&P 500 and member's present rating.

4

Lastly, an overall probability of attrition was determined. Because the data only contains Sailors who experienced events within the years 2001 to 2008, we used conditional probabilities in order to calculate the attrition rate. The research question was: What is the likelihood that given an event, the event is an attrite? Mathematically this is P (Attrite $|$ Event). The result was that the attrition rate during the examined period was 36 percent for Sailors at or less than four years of service. This was calculated by P (.83857 $|$ .43186) = .362144 or 36 percent attrition.

Conclusions and Recommendations

Predicting the attrition rate of the Total Force and the ability to predict attrition at the individual level are two strong outcomes of this initial step in predictive modeling research that will benefit the U. S. Navy. The results yielded from the research demonstrate that it is possible to very accurately predict attrition based on individual characteristics. If the model was deployed using real-time data, the Navy could potentially experience cost savings in part due to better information about Sailors who have a high likelihood of being an attrite and estimated time periods in which the attrition will occur. Retention strategies could be employed at the individual level well in advance of attriting in order to minimize this occurrence.

Possible future research could include larger data sets, and including more exogenous economic variables, as well as examining different groups rather than just first-term or enlisted Sailors.

It is recommended that groups within the U. S. Navy such as Manpower, Personnel, Training and Education (N1); Navy Recruiting Command (NRC); and others look at employing data mining methods and models in order to better predict attrition and capitalize on retaining top performing Sailors.

References

Buddin, R. (1984). *Analysis of Early Military Attrition Behavior*. Santa Monica, CA: The Rand Corporation.

Buttrey, S. E., & Larson, H. (1999). *Determining Characteristic Groups to Predict Army Attrition*. Navy Postgraduate School.

Carlson, C. G. (1981). *A Descriptive Analysis of First Term Attrition from U. S. Naval Ships*. Navy Postgraduate School.

Cox, D. R. (1972). Regression models and life tables. *Journal of Royal Statistical Society Series B; 34*: 187–220.

Flyer, E. S. (1959). *Factors Relating to Discharge for Unsuitability Among 1956 Airman Accessions to the Air Force*. Lackland AFB Personnel Research Laboratory.

Guidici, P. (2003). *Applied Data Mining: Statistical Methods for Business and Industry*. England: Wiley.

Hosek, J. R., Antel, J., & Peterson, C. E. (1989). Who Stays, Who Leaves? Attrition Among First Term Enlistees. *Armed Forces and Society, 15(3)*, 389–409.

McCloy, R. A., DiFazio, A. S., & Carter, G. W. (1993). *Prediction of First-Term Military Attrition Using Pre-Enlistment Predictors*. Annual Meeting of the Physcological Association. Toronto, Ontario, Canada.

Salo, M., & Siebold, G. (2006). *Predictors of Attrition in the Finnish Conscript Service*. United States Army Research Institute for the Behavioral and Social Sciences.

White, M.A. , Mottern, J.A. et al (2007), *First Watch on the First Term of Enlistment: Cross-Sectional and Longitudinal Analysis of Data from the First Year of the Study.* Naval Personnel, Research, Studies, and Technology.
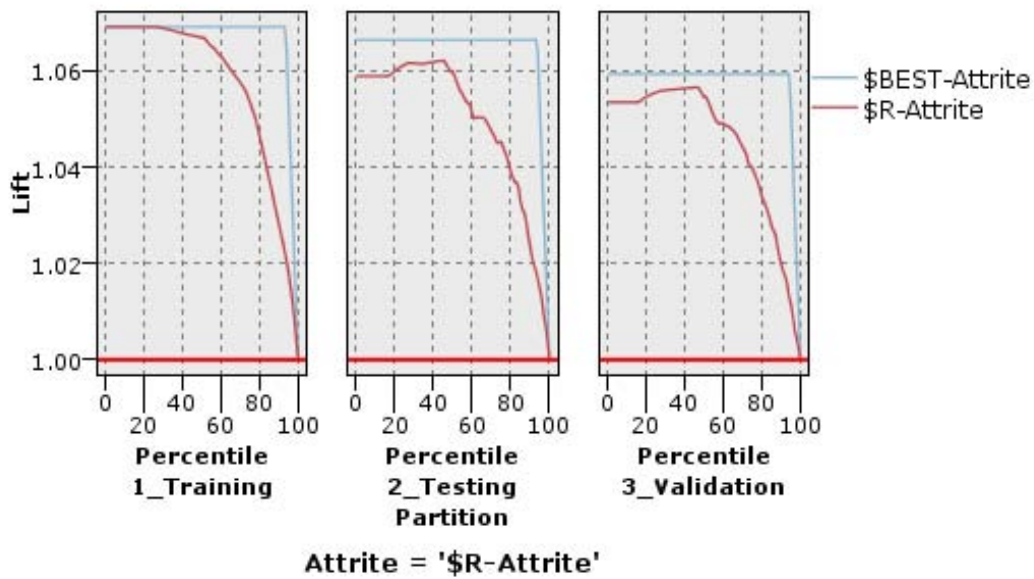
Appendix A:
Lift Charts

Figure A-1. CHAID algorithm lift chart.

Figure A-1 shows a lift curve in decreasing scoring for each decile. The Blue line is the best possible model, with the red lines indicating the model's performance. In the training partition above, the graph shows that using this model one is 1.06 times more likely to get a success (predict attrition).
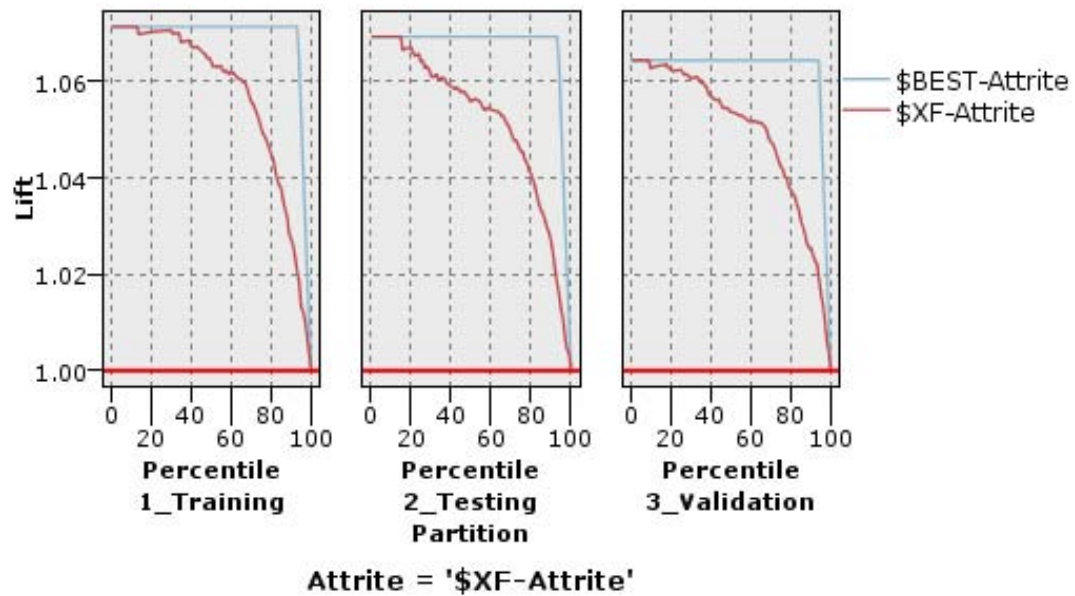
Figure A-2. ENSEMBLE algorithm lift chart.

Again Figure A-2 shows a lift curve in decreasing scoring for each decile. The Blue line is the best possible model, with the red lines indicating the model's performance. In the training partition above, the graph shows that using the model one is 1.06 times more likely to get a success (predict attrition) than with out it. However, the ensemble gets progressively worse as one moves to the validation partition if compared with the previous CHAID. Therefore, the CHAID is a better model in this instance.

Appendix B:
Model Misclassification Rates

Table B-1 is sometimes referred to as a misclassification chart or confusion matrix and shows overall accuracy. If an observation actually was an attrite, then the model predicted it to be an attrite 59 percent of the time, and predicted events as non-events (Type I error or False negative) 40.537 or 41 percent. However, if a record was not an attrite then the model said it was not 99 percent of the time with a Type II error of less than 1 percent.

<div align="center">

Table B-1
CHAID outcome

Attrite	0	1
0	99.160	.840
1	40.537	59.463

</div>

Table B-2 shows overall accuracy. If an observation actually was an attrite, then the model predicted it to be an attrite 57 percent of the time, and predicted events as non-events (Type I error or False negative) 42.879 or 43 percent. However, if a record was not an attrite then the model said it was not 99 percent of the time with a Type II error of less than 1 percent.

<div align="center">

Table B-2
ENSEMBLE outcome

Attrite	0	1
0	99.020	.980
1	42.879	57.121

</div>

Distribution

AIR UNIVERSITY LIBRARY
ARMY RESEARCH INSTITUTE LIBRARY
ARMY WAR COLLEGE LIBRARY
CENTER FOR NAVAL ANALYSES LIBRARY
HUMAN RESOURCES DIRECTORATE TECHNICAL LIBRARY
JOINT FORCES STAFF COLLEGE LIBRARY
MARINE CORPS UNIVERSITY LIBRARIES
NATIONAL DEFENSE UNIVERSITY LIBRARY
NAVAL HEALTH RESEARCH CENTER WILKINS BIOMEDICAL LIBRARY
NAVAL POSTGRADUATE SCHOOL DUDLEY KNOX LIBRARY
NAVAL RESEARCH LABORATORY RUTH HOOKER RESEARCH LIBRARY
NAVAL WAR COLLEGE LIBRARY
NAVY PERSONNEL RESEARCH, STUDIES, AND TECHNOLOGY SPISHOCK
    LIBRARY (3)
OFFICE OF NAVAL RESEARCH (CODE 34)
PENTAGON LIBRARY
USAF ACADEMY LIBRARY
US COAST GUARD ACADEMY LIBRARY
US MERCHANT MARINE ACADEMY BLAND LIBRARY
US MILITARY ACADEMY AT WEST POINT LIBRARY
US NAVAL ACADEMY NIMITZ LIBRARY